

Prediction-Based Task Assignment on Spatial Crowdsourcing

Peng Cheng[#], Xiang Lian^{*}, Lei Chen[#], Cyrus Shahabi[†]

[#]Hong Kong University of Science and Technology, Hong Kong, China
 {pchengaa, leichen}@cse.ust.hk

^{*}Kent State University, Ohio, USA
 xlian@kent.edu

[†]University of Southern California, California, USA
 shahabi@usc.edu

Abstract—With the rapid development of mobile devices and crowdsourcing platforms, the *spatial crowdsourcing* has recently attracted much attention from the database community. Specifically, the spatial crowdsourcing refers to a system that periodically assigns a number of location-based workers with spatial tasks nearby (e.g., taking photos or videos at some spatial locations). Previous works on the spatial crowdsourcing usually designed task assignment strategies that maximize some assignment scores, which are however only based on available workers/tasks in the system at the time point of assigning workers/tasks. These strategies may achieve local optimality, due to the neglect of future workers/tasks that may join the system. In contrast, in this paper, we aim to achieve “globally optimal” task assignments, by considering not only those present, but also future (via predictions), workers/tasks. Specifically, we formalize an important problem, namely *prediction-based spatial crowdsourcing* (PB-SC), which expects to obtain a better local optimal strategy for worker-and-task assignments, over both present and predicted task/worker locations, such that the total assignment quality score is maximized (for both current and future timestamps), under the constraint of the traveling budget.

The PB-SC problem is quite challenging, in terms of the prediction accuracy and efficiency. In this paper, we design an effective grid-based prediction method to estimate spatial distributions of workers/tasks in the future, and then utilize the predicted ones in our procedure of worker-and-task assignments. We prove that, the PB-SC problem is NP-hard, and thus intractable. Therefore, we propose efficient approximate algorithms to tackle the PB-SC problem, including *greedy* and *divide-and-conquer* (D&C) approaches, which can efficiently assign workers to spatial tasks with high quality scores and low budget consumptions, by considering both current and future task/worker distributions. Through extensive experiments, we demonstrate the efficiency and effectiveness of our PB-SC processing approaches on real/synthetic data.

I. INTRODUCTION

The popularity of smart devices not only brings the convenience to our daily life, but also enables people to easily participate in some location-based tasks nearby, such as taking photos/videos (e.g., street view of Google Maps [2]), monitoring traffic conditions (e.g., Waze [4]), and checking display shelves at neighborhood stores (e.g., Gigwalk [1]). Recently, to exploit these phenomena, a new framework, namely *spatial crowdsourcing* [18], for requesting workers to conduct spatial tasks, has drawn much attention from both academia (e.g., MediaQ [19]) and industry (e.g., TaskRabbit [3]). A typical spatial crowdsourcing system (e.g., gMission

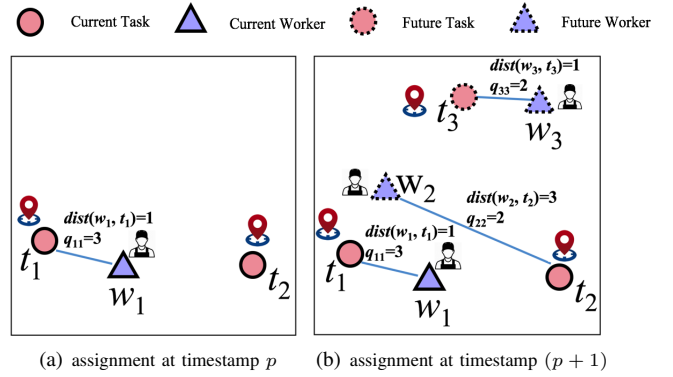


Fig. 1. Locally Optimal Worker-and-Task Assignments in the Spatial Crowdsourcing System.

[8]) distributes a number of dynamically moving workers to accomplish spatial tasks (e.g., taking photos/videos), which requires workers to physically go to the specified locations to complete these tasks.

We first give the following motivation example.

Example 1 (The Spatial Crowdsourcing Problem with Multiple Rounds) Consider a scenario of the spatial crowdsourcing in Figure 1, where spatial tasks, $t_1 \sim t_3$, are represented by red circles, and workers, $w_1 \sim w_3$, are denoted by blue triangles. In particular, Figure 1(a) shows a worker, w_1 , and two tasks, t_1 and t_2 , which join the system in the first round at a timestamp p (denoted by markers with solid border). Figure 1(b) depicts two more workers, w_2 and w_3 , and one more task, t_3 , which arrive at the system in the next (second) round at a future timestamp $(p+1)$ (denoted by markers with the dashed border).

Each worker w_i ($1 \leq i \leq 3$) has one’s own expertise to perform different types of spatial tasks t_j ($1 \leq j \leq 3$). Thus, we assume that the competence of a task t_j by worker w_i can be captured by a quality score q_{ij} (as shown in Table I). Furthermore, each worker, w_i , is provided with some salary (e.g., allowance, award, or credits) to cover the traveling cost from w_i to t_j , which is proportional to the traveling distance, $\text{dist}(w_i, t_j)$ (as depicted in Table I), where $\text{dist}(x, y)$ is a distance function from x to y . This way, workers, w_i , are encouraged to do tasks, t_j , even if they are far away from each other, which is useful for enhancing the task completion rate, especially for workers/tasks with unbalanced location distributions.

Given a maximum salary budget in each round, an

worker-and-task pair, $\langle w_i, t_j \rangle$	distance, $dist(w_i, t_j)$	quality score, q_{ij}
$\langle w_1, t_1 \rangle$	1	3
$\langle w_1, t_2 \rangle$	2	2
$\langle w_1, t_3 \rangle$	4	2
$\langle w_2, t_1 \rangle$	1	4
$\langle w_2, t_2 \rangle$	3	2
$\langle w_2, t_3 \rangle$	2	1
$\langle w_3, t_1 \rangle$	5	2
$\langle w_3, t_2 \rangle$	3	1
$\langle w_3, t_3 \rangle$	1	2

TABLE I. DISTANCES AND QUALITY SCORES OF WORKER-AND-TASK PAIRS

important spatial crowdsourcing problem is to assign workers w_i to tasks t_j at both current and future timestamps, p and $(p + 1)$, respectively, such that the total quality score of assignments is maximized, and the total salary for workers is under the budget constraint.

In Example 1, traditional spatial crowdsourcing approaches [10], [18], [19] usually considered the worker-and-task assignments in separate rounds, only based on workers/tasks that are available in the system for each round. For example, in the first round of Figure 1(a), only t_1 , t_2 , and w_1 exist in the system at timestamp p . Therefore, we assign worker w_1 to task t_1 (rather than task t_2), since it holds that $dist(w_1, t_1) < dist(w_1, t_2)$ and $q_{11} > q_{12}$ (i.e., worker w_1 can accomplish task t_1 with lower budget consumption and higher quality, compared with task t_2 , as given in Table I). Next, in the second round of Figure 1(b), 2 tasks, t_2 and t_3 , and 2 workers, w_2 and w_3 , are available in the system at timestamp $(p+1)$. Traditional spatial crowdsourcing approaches would obtain assignment pairs $\langle w_2, t_2 \rangle$ and $\langle w_3, t_3 \rangle$ in this round (see lines in Figure 1(b)). As a result, such an assignment strategy in 2 separate rounds leads to the total traveling cost 5 ($= 1 + 3 + 1$) and total quality score 7 ($= 3 + 2 + 2$).

Note that, the assignment strategy above does not take into account future workers/tasks that may join the system in a later round. This may result in a local optimality (from the view of assignments for multiple rounds over a long period of time). This is because, at timestamp p , if we somewhat know future workers/tasks that arrive at timestamp $(p+1)$, then we may obtain a better assignment strategy (i.e., with lower traveling cost and higher quality score). Based on this observation, in this paper, we will propose a novel and useful spatial crowdsourcing problem, namely *prediction-based spatial crowdsourcing* (PB-SC), which assigns moving workers to spatial tasks (with their current/predicted location and quality distributions), with the highest reliability, under the budget constraint, and with the expectation of achieving global assignment optimality.

Example 2 (The Prediction-Based Spatial Crowdsourcing Problem) In the example of Figure 1, the PB-SC problem is to obtain/predict the locations/qualities of workers/tasks in both present and future rounds (i.e., at timestamps p and $(p+1)$, respectively), such that the total quality score of assignments is maximized, and the total traveling cost is under budget constraints. As shown in Figure 2, at timestamp p , we can have the prediction-based assignments: $\langle w_2, t_1 \rangle$, $\langle w_1, t_2 \rangle$, and $\langle w_3, t_3 \rangle$, which can achieve global optimality with smaller traveling cost 4 ($= 1 + 2 + 1$) and higher quality score 8 ($= 4 + 2 + 2$), compared to locally optimal assignments (without prediction) in Figure 1 with the traveling cost, 5, and the quality score 7, respectively.

From the examples above, we can see that the worker-

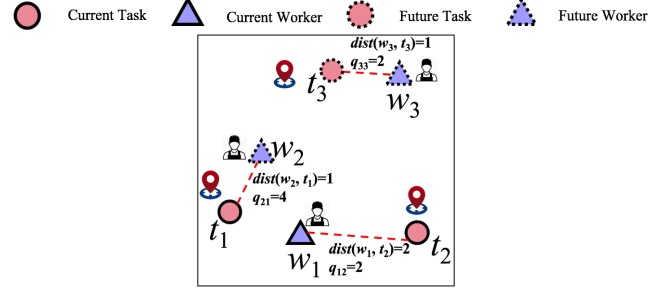


Fig. 2. Globally Optimal Assignments for the Prediction-Based Spatial Crowdsourcing (PB-SC) Problem.

and-task assignment only based on the currently available tasks/workers (i.e., without prediction) in each round may lead to local optimal solution (e.g., with high traveling cost and low quality score). In contrast, this paper studies the PB-SC problem, which aims to accurately predict future location distributions of new tasks/workers, and efficiently achieve a better assignment strategy (with high quality score and under budget constraints) over both existing and predicted tasks/workers.

Different from prior works on the spatial crowdsourcing problem over current workers/tasks, our PB-SC problem requires designing an accurate prediction approach for estimating future location distributions of tasks and workers (and the quality distributions of worker-and-task assignment pairs as well), and considering the assignments over the estimated location/quality variables of future tasks/workers (rather than deterministic location/quality values), which is quite challenging. Furthermore, in this paper, we prove that, our PB-SC problem is NP-hard, by reducing it from the 0-1 Knapsack problem [22]. As a result, the PB-SC problem is not tractable. Therefore, in order to efficiently tackle the PB-SC problem, we will propose effective approximate algorithms, including *greedy* and *divide-and-conquer* (D&C) approaches, over both current and predicted workers/tasks, which can efficiently compute sub-optimal assignment pairs with high quality score under the budget constraint.

Specifically, we make the following contributions.

- We formally define the *prediction-based spatial crowdsourcing* (PB-SC) problem in Section II, over both current and predicted worker/task location/quality distributions in the spatial crowdsourcing system. We prove that the PB-SC problem is NP-hard in Section II-D.
- We propose an effective grid-based prediction approach to estimate location/quality distributions/statistics for future tasks/workers in Section III.
- We design an efficient *greedy algorithm* to iteratively select one best assignment pair each time over current/future tasks/workers in Section IV.
- We illustrate a novel *divide-and-conquer algorithm* (D&C) to recursively divide the problem into subproblems and merge the assignment results in subproblems in Section V.
- We verify the effectiveness and efficiency of our proposed PB-SC approaches with extensive experiments on real and synthetic data sets in Section VI.

In addition to the contributions listed above, in this paper, we review previous works on spatial crowdsourcing in Section VII, and conclude in Section VIII.

II. PROBLEM DEFINITION

A. Dynamically Moving Workers

We first define the moving workers in the spatial crowdsourcing system.

Definition 1: (Dynamically Moving Workers) Let $W_p = \{w_1, w_2, \dots, w_n\}$ be a set of n moving workers at timestamp p . Each worker w_i ($1 \leq i \leq n$) is located at position $l_i(p)$ at timestamp p , and freely moves with velocity v_i .

In Definition 1, workers w_i can dynamically move with speed v_i in any direction. At each timestamp p , workers w_i are located at positions $l_i(p)$. They can freely join or leave the spatial crowdsourcing system.

B. Time-Constrained Spatial Tasks

We define spatial tasks in the spatial crowdsourcing system, which are constrained by deadlines of arriving at task locations.

Definition 2: (Time-Constrained Spatial Tasks) Let $T_p = \{t_1, t_2, \dots, t_m\}$ be a set of time-constrained spatial tasks at timestamp p . Each spatial task t_j ($1 \leq j \leq m$) is located at a specific location l_j , and workers are expected to reach the location of task t_j before the deadline e_j .

In Definition 2, a task requester creates a time-constrained spatial task t_j (e.g., taking a photo), which requires workers physically moving to a specific location l_j , and arriving at l_j before the deadline e_j . Note that, in this paper, we assume that each spatial task can be accomplished by one worker.

C. The Prediction-Based Spatial Crowdsourcing Problem

In this subsection, we will formalize our problem of the *prediction-based spatial crowdsourcing* (PB-SC), which assigns time-constrained spatial tasks to spatially scattered workers, such that the total quality score of assignments can be maximized with global optimality (defined as optimal assignments within a time interval P), under the traveling budget constraint.

Task Assignment Instance Set. Before we present the PB-SC problem, we first introduce the notion of the *task assignment instance set*.

Definition 3: (Task Assignment Instance Set, I_p) At timestamp p , given a worker set W_p and a task set T_p , a *task assignment instance set*, denoted by I_p , is a set of valid worker-and-task assignment pairs in the form $\langle w_i, t_j \rangle$, where each worker $w_i \in W_p$ is assigned to at most one task $t_j \in T_p$, and each task $t_j \in T_p$ is accomplished by at most one worker $w_i \in W_p$.

Intuitively, I_p in Definition 3 is one possible (valid) worker-and-task assignment between worker set W_p and task set T_p . A valid assignment pair $\langle w_i, t_j \rangle$ is in I_p , if and only if this pair satisfies the conditions that, worker w_i can reach the location l_j of task t_j before the deadline e_j .

The Salary of the Worker-and-Task Assignment Pair. As discussed earlier in Section I, we assume that each worker-and-task assignment pair $\langle w_i, t_j \rangle$ is associated with a traveling cost, c_{ij} , which corresponds to the salary (e.g., allowance, award, or system credits) to cover the transportation fee from $l_i(p)$ to l_j (e.g., cost of gas or public transportation). Formally, the salary, c_{ij} , is given by: $c_{ij} = C \cdot \text{dist}(l_i(p), l_j)$, where C is the unit cost per mile, and $\text{dist}(l_i(p), l_j)$ is the distance between locations of worker w_i and task t_j . For simplicity, we will use the Euclidean distance function $\text{dist}(l_i(p), l_j)$ in

this paper. We leave the topics of using other distance metrics (e.g., road-network distances) as our future work.

The Quality Score of the Worker-and-Task Assignment Pair. Each worker-and-task assignment pair $\langle w_i, t_j \rangle$ is also associated with a *quality score*, denoted as q_{ij} ($\in [0, 1]$), which indicates the quality of the task t_j that is completed by worker w_i . Intuitively, q_{ij} is a score to measure the quality that a task t_j can be finished by w_i .

Here, due to different types of tasks with distinct difficulty levels (e.g., simply taking photos, or repairing houses in a complex way) and different expertise or competence of workers in performing tasks (e.g., with different backgrounds or skills), we assume that different worker-and-task pairs $\langle w_i, t_j \rangle$ may be associated with diverse quality scores q_{ij} .

The PB-SC Problem. Next, we give the definition of our *prediction-based spatial crowdsourcing* (PB-SC) problem.

Definition 4: (Prediction-based Spatial Crowdsourcing Problem) Given a time interval P and a maximum salary budget B for each round, the problem of *prediction-based spatial crowdsourcing* (PB-SC) is to assign the available workers $w_i \in W_p$ to spatial tasks $t_j \in T_p$ to obtain a task assignment instance set, I_p , for each round at timestamp $p \in P$, such that:

- 1) for any round at timestamp $p \in P$, each worker $w_i \in W_p$ is assigned to at most one spatial task $t_j \in T_p$ such that his/her arrival time at location l_j is before deadline e_j ;
- 2) for any round at timestamp $p \in P$, the total salary (i.e., the traveling cost) of all the assigned workers does not exceed budget B , that is, $\sum_{\langle w_i, t_j \rangle \in I_p} c_{ij} \leq B$; and
- 3) the total quality score of the assigned tasks for all rounds (at timestamps $p \in P$) is maximized, that is,

$$\text{maximize } \sum_{p \in P} \sum_{\langle w_i, t_j \rangle \in I_p} q_{ij}. \quad (1)$$

Intuitively, the PB-SC problem (given in Definition 4) assigns workers to tasks in multiple rounds within a time interval P , such that (1) for each round, assignment pairs $\langle w_i, t_j \rangle$ are valid (i.e., satisfying the time constraints, e_j); (2) in each round, the total salary of assignment pairs is lower than or equal to the maximum budget B ; and (3) the total quality score of assignments for all rounds is maximized.

As discussed earlier in Section I, it is not globally optimal (i.e., with maximal total quality score of optimal assignments within the time interval P) to simply consider the assignments over available tasks/workers in the system in multiple rounds separately. The PB-SC problem aims to compute a globally optimal assignment solution, by taking into account tasks/workers not only at the current timestamp p , but also at the future timestamp $(p + 1)$ in the next round (i.e., $(p + 1) \in P$).

Table II summarizes the commonly used symbols.

D. Hardness of the PB-SC Problem

With n dynamically moving workers and m time-constrained spatial tasks, in the worst case, there are an exponential number of possible worker-and-task assignment strategies, which leads to high time complexity (i.e., $O((m + 1)^n)$). In this subsection, we prove that our PB-SC problem is NP-hard, by reducing it from a well-known NP-hard problem, *0-1 Knapsack problem* [22].

TABLE II. SYMBOLS AND DESCRIPTIONS.

Symbol	Description
T_p	a set of m time-constrained spatial tasks t_j at timestamp p
W_p	a set of n dynamically moving workers w_i at timestamp p
\hat{w}_i (or \hat{t}_j)	the predicted worker (or task)
\bar{w}_i (or \bar{t}_j)	the worker (or task) in either current or next round
I_p	the task assignment instance set at timestamp p
e_j	the deadline of arriving at the location of task t_j
$l_i(p)$	the position of worker w_i at timestamp p
l_j	the position of task t_j
v_i	the velocity of worker w_i
c_{ij}	the traveling cost from the location of worker w_i to that of task t_j
q_{ij}	the quality score of assigning worker w_i to conduct task t_j
B	the maximum salary budget (i.e., traveling cost) in each round
C	the unit price of the traveling cost by workers
P	a time interval
w	the size of the sliding window to do the prediction

Lemma 2.1: (Hardness of the PB-SC Problem) The prediction-based spatial crowdsourcing (PB-SC) problem is NP-hard.

Proof: Please refer to Appendix A of our technical report [9]. ■

From Lemma 2.1, we can see that the PB-SC problem is NP-hard, and thus intractable. Alternatively, we will later design efficient approximate algorithms to tackle the PB-SC problem and achieve sub-optimal solutions.

E. Framework

Figure 3 illustrates a general framework, namely procedure PB-SC_Framework, for solving the PB-SC problem, which assigns workers with spatial tasks for multiple rounds (within a time interval P), based on the predicted location/quality distributions of workers/tasks.

Specifically, for each round at timestamp p , we first retrieve a set, T_p , of available spatial tasks, and a set, W_p , of available workers (lines 1-3). Here, the task set T_p (or worker set W_p) contains both tasks (or workers) that have not been assigned in the last round and the ones that newly arrive at the system after the last round (note: those workers who finished tasks in previous rounds are also treated as “new workers” that join the system). In order to avoid the local optimality at the current timestamp p , we also need to predict location/quality distributions of workers and tasks at a future timestamp $(p+1)$ that newly join the system, and obtain two sets W_{p+1} and T_{p+1} , respectively (line 4).

Then, with both current and future sets of tasks/workers (i.e., T_p/W_p and T_{p+1}/W_{p+1} , respectively), we can apply our proposed algorithms in this paper (including *greedy* and *divide-and-conquer* (D&C)), and retrieve better local optimal assignment pairs in assignment instance set I_p (line 5). For each pair $\langle w_i, t_j \rangle$ in I_p , we notify worker w_i to do task t_j (lines 6-7).

Procedure PB-SC_Framework {

Input: a time interval P

Output: a worker-and-task assignment strategy within the time interval P

- (1) **for** each round at timestamp $p \in P$
- (2) retrieve all the available spatial tasks in set T_p
- (3) retrieve all the available workers in set W_p
- (4) predict new future tasks/workers in T_{p+1} and W_{p+1} in the next round
- (5) apply the *greedy* or *divide-and-conquer* approach to obtain a assignment instance set, I_p , w.r.t. T_p , W_p , T_{p+1} and W_{p+1}
- (6) **for** each pair $\langle w_i, t_j \rangle$ in I_p
- (7) inform worker w_i to conduct task t_j

Fig. 3. A Framework for Tackling the PB-SC Problem.

III. THE GRID-BASED WORKER/TASK PREDICTION APPROACH

In order to achieve the global optimality in our PB-SC problem, we need to accurately predict the future scenario of workers/tasks that newly join the spatial crowdsourcing

system. Specifically, in this section, we will introduce a grid-based worker/task prediction approach, which can efficiently and effectively estimate the number of future workers/tasks, location distributions of future workers/tasks, and quality score distributions w.r.t. future worker-and-task pairs (and their existence probabilities as well).

A. The Grid-Based Prediction Algorithm

In this subsection, we discuss how to predict the number of tasks /workers, and their location distributions in a 2-dimensional data space $\mathcal{U} = [0, 1]^2$. In particular, we consider a grid index, \mathcal{I} , over tasks and workers, which divides the data space \mathcal{U} into γ^2 cells, each with the side length $1/\gamma$, where the selection of the best γ value can be guided by a cost model in [10]. Then, we will estimate possible workers/tasks that may fall into each cell at a future timestamp, which can be inferred from historical data within the most recent sliding window of size w .

In particular, our grid-based prediction algorithm first predicts the future numbers of workers/tasks from the latest w worker/task sets in cells, then generates possible worker (or task) samples in W_{p+1} (or T_{p+1}) for cells of the grid index \mathcal{I} .

First, we initialize a worker set W_{p+1} and a task set T_{p+1} in the future round with empty sets. Then, within each cell $cell_i$, we can obtain its w latest worker counts, $|W_{p-w+1}^{(i)}|$, $|W_{p-w+2}^{(i)}|$, ..., and $|W_p^{(i)}|$, which form a *sliding window* of a time series (with size w). Due to the temporal correlation of worker counts in the sliding window, in this paper, we utilize the *linear regression* [20] over these w worker counts to predict the future number, $|W_{p+1}^{(i)}|$, of workers in cell, $cell_i$, that newly join the system at timestamp $(p+1)$. Note that, other prediction methods can be also plugged into our grid-based prediction framework, and we would like to leave it as our future work. Similarly, we can estimate the number, $|T_{p+1}^{(i)}|$, of tasks in $cell_i$ for the next round at timestamp $(p+1)$.

According to the predicted numbers of workers/tasks, we can uniformly generate $|W_{p+1}^{(i)}|$ worker samples (or $|T_{p+1}^{(i)}|$ task samples) within each cell $cell_i$, and add them to the predicted worker set W_{p+1} (or task set T_{p+1}). Finally, we return the two sets W_{p+1} and T_{p+1} . For the pseudo code of the PB-SC prediction algorithm, please refer to Appendix B of our technical report [9].

Location Distributions of the Predicted Workers/Tasks.

Note that, random samples of workers/tasks in cells can approximately capture future location distributions of workers/tasks. However, in each cell, discrete samples may be of small sample sizes, which may lead to low prediction accuracy. For example, if the sample size is only 1, then different possible locations of this one single sample (generated in the cell) may dramatically affect our PB-SC assignment results, with such a predicted sample.

Inspired by this, instead of using discrete samples predicted, in this paper, we will alternatively consider continuous *probability density function* (pdf) for location distributions of workers/tasks. Specifically, we apply the *kernel density estimation* over samples in each cell to describe the distributions of samples' locations. That is, centered at each sample s_i generated in $cell_i$, we can obtain the continuous pdf function of this worker/task sample, $f(x) = \prod_{r=1}^2 \left(\frac{1}{h_r} K \left(\frac{x[r]-s[r]}{h_r} \right) \right)$, where $h_r \in (0, 1)$ is

the bandwidth on the r -th dimension, and function $K(\cdot)$ is a *uniform kernel function* [15], given by $K(u) = \frac{1}{2} \cdot \mathbf{1}(|u| \leq 1)$. Here, $\mathbf{1}(|u| \leq 1) = 1$, when $|u| \leq 1$ holds. Note that, the choice of the kernel function is not significant for approximation results [15], thus, in this paper, we use uniform kernel function $K(\cdot)$. From the pdf function $f(x)$ of each sample $s_i \in \text{cell}_i$, each dimension r can be bounded by an interval $[s_i[r] - h_r, s_i[r] + h_r]$.

Typically, in the literature [15], we set $h_r = \hat{\sigma} C_v(k) n^{-1/(2v+1)}$, where $\hat{\sigma}$ is the standard deviation of samples (derived from current worker/task statistics), v is the order of the kernel (v is set to 2 here), and $C_v(k) = 1.8431$ ($= 2 \left(\frac{\pi^{1/2} (v!)^3 R(k)}{2v(2v-1)k_v^2(k)} \right)^{1/(2v+1)}$, for $k_v(k) = 1/3, R(k) = 1/2$ with Uniform kernel functions).

B. Statistics of the Predicted Workers/Tasks

For the ease of the presentation, in this paper, we denote those future workers w_i and tasks t_j that are predicted as \widehat{w}_i and \widehat{t}_j , respectively.

Due to the predicted future workers/tasks, in our PB-SC problem, we need to consider worker-and-task assignment pairs that may involve predicted workers/tasks. That is, we have 3 cases, $\langle \widehat{w}_i, t_j \rangle$, $\langle w_i, \widehat{t}_j \rangle$, and $\langle \widehat{w}_i, \widehat{t}_j \rangle$, where \widehat{w}_i and \widehat{t}_j are the predicted worker and task samples, respectively, following uniform distributions represented by kernel functions $K(\cdot)$ (as mentioned in Section III-A).

Due to the existence of these predicted workers/tasks, the traveling costs and the quality scores of assignment pairs now become random variables (rather than fixed values). In this subsection, we will discuss how to obtain statistics (e.g., mean and variance) of traveling costs, quality scores, and confidences, associated with assignment pairs.

The Traveling Cost of Pairs with the Predicted Workers/Tasks. The traveling cost, c_{ij} , of worker-and-task pairs involving the predicted workers/tasks (i.e., $\langle \widehat{w}_i, t_j \rangle$, $\langle w_i, \widehat{t}_j \rangle$, or $\langle \widehat{w}_i, \widehat{t}_j \rangle$) can be given by $C \cdot \text{dist}(\widehat{w}_i, t_j)$, $C \cdot \text{dist}(w_i, \widehat{t}_j)$, or $C \cdot \text{dist}(\widehat{w}_i, \widehat{t}_j)$, respectively.

We discuss the general case of computing statistics of variable $\widehat{c}_{ij} = C \cdot \text{dist}(\widehat{w}_i, \widehat{t}_j)$. Since it is nontrivial to calculate the statistics of the Euclidean distance between variables \widehat{w}_i and \widehat{t}_j , we alternatively consider statistics (mean and variance) of the squared Euclidean distance variable $Z^2 = \text{dist}^2(\widehat{w}_i, \widehat{t}_j)$ ($= \sum_{r=1}^2 (\widehat{w}_i[r] - \widehat{t}_j[r])^2$), where \widehat{w}_i and \widehat{t}_j are two variables uniformly residing in a 2D space.

The Computation of Mean $E(Z^2)$. Let variable $Z_r = \widehat{w}_i[r] - \widehat{t}_j[r]$, for $r = 1, 2$, whose mean $E(Z_r)$ and variance $\text{Var}(Z_r)$ can be easily computed (i.e., $E(Z_r) = s_i[r] - s_j[r]$ and $\text{Var}(Z_r) = \frac{(h_r(\widehat{w}_i[r]))^2 + (h_r(\widehat{t}_j[r]))^2}{3}$ respectively).

Then, we have $Z^2 = Z_1^2 + Z_2^2$. Thus, the mean $E(Z^2)$ can be given by:

$$E(Z^2) = E(Z_1^2) + E(Z_2^2). \quad (2)$$

The Computation of Variance $\text{Var}(Z^2)$. Moreover, for variance $\text{Var}(Z^2)$, it holds that:

$$\begin{aligned} \text{Var}(Z^2) &= E(Z^4) - (E(Z^2))^2 \\ &= E((Z_1^2 + Z_2^2)^2) - (E(Z^2))^2 \\ &= E(Z_1^4) + 2 \cdot E(Z_1^2) \cdot E(Z_2^2) + E(Z_2^4) - (E(Z^2))^2. \end{aligned} \quad (3)$$

From Eqs. (2) and (3) above, the remaining issues are to compute $E(Z_r^2)$ and $E(Z_r^4)$ (for $r = 1, 2$).

The Computation of $E(Z_r^2)$. For $E(Z_r^2)$, since $Z_r = \widehat{w}_i[r] - \widehat{t}_j[r]$, we have:

$$\begin{aligned} E(Z_r^2) &= \text{Var}(Z_r) + (E(Z_r))^2 \\ &= \text{Var}(\widehat{w}_i[r]) + \text{Var}(\widehat{t}_j[r]) + (E(\widehat{w}_i[r]) - E(\widehat{t}_j[r]))^2. \end{aligned} \quad (4)$$

The Computation of $E(Z_r^4)$. For $E(Z_r^4)$, we can derive that:

$$\begin{aligned} E(Z_r^4) &= E((\widehat{w}_i[r] - \widehat{t}_j[r])^4) \\ &= E(\widehat{w}_i[r]^4) - 4 \cdot E(\widehat{w}_i[r]^3) \cdot E(\widehat{t}_j[r]) \\ &\quad + 6 \cdot E(\widehat{w}_i[r]^2) \cdot E(\widehat{t}_j[r]^2) - 4 \cdot E(\widehat{w}_i[r]) \cdot E(\widehat{t}_j[r]^3) \\ &\quad + E(\widehat{t}_j[r]^4). \end{aligned} \quad (5)$$

In Eq. (5), variable $\widehat{w}_i[r]$ follows the uniform distribution within bound $[lb_w, ub_w]$ (for the r -th dimension of uniform kernel function $K(\cdot)$ in Section III-A). We can thus infer that:

$$\begin{aligned} E(\widehat{w}_i[r]^4) &= \int_{lb_w}^{ub_w} x^4 \frac{1}{ub_w - lb_w} dx = \frac{ub_w^5 - lb_w^5}{5(ub_w - lb_w)}, \\ E(\widehat{w}_i[r]^3) &= \int_{lb_w}^{ub_w} x^3 \frac{1}{ub_w - lb_w} dx = \frac{ub_w^4 - lb_w^4}{4(ub_w - lb_w)}, \\ E(\widehat{w}_i[r]^2) &= \int_{lb_w}^{ub_w} x^2 \frac{1}{ub_w - lb_w} dx = \frac{ub_w^3 - lb_w^3}{3(ub_w - lb_w)}, \end{aligned}$$

where $[lb_w, ub_w] = [s_i[r] - h_r(\widehat{w}_i[r]), s_i[r] + h_r(\widehat{w}_i[r])]$.

Similarly, we can also obtain $E(\widehat{t}_j[r]^4)$, $E(\widehat{t}_j[r]^3)$, and $E(\widehat{t}_j[r]^2)$ for task $\widehat{t}_j[r]$. We omit it here.

This way, by substituting Eqs. (4) and (5) into Eqs. (2) and (3), we can obtain mean $E(Z^2)$ and variance $\text{Var}(Z^2)$ of the squared Euclidean distance between two Uniform distributions.

Quality Scores of Pairs with the Predicted Workers/Tasks. We consider the three cases to compute statistics of quality score distributions.

Case 1: $\langle \widehat{w}_i, t_j \rangle$. In this case, in the current round at timestamp p , we can obtain all the n_i workers w_i that can reach task t_j . Then, we use quality scores, q_{ij} , of their corresponding worker-and-task pairs $\langle w_i, t_j \rangle$ as samples (each with probability $1/n_i$), which can describe/estimate future distributions of quality scores. Correspondingly, with these samples, we can obtain mean and variance of quality scores between the predicted worker \widehat{w}_i and the current task t_j .

Case 2: $\langle w_i, \widehat{t}_j \rangle$. Similar to Case 1, we can obtain m_j spatial tasks t_j that can be reached by worker w_i . Then, we obtain m_j quality score samples from valid pairs $\langle w_i, t_j \rangle$ (each sample with probability $1/m_j$), whose mean and variance can be used to capture the quality score distribution between the current worker w_i and a predicted task \widehat{t}_j .

Case 3: $\langle \widehat{w}_i, \widehat{t}_j \rangle$. Since both worker \widehat{w}_i and task \widehat{t}_j have predicted distributions, we cannot directly obtain quality score distributions. Thus, our basic idea is to infer future quality scores by existing workers w_i and tasks t_j at the current timestamp p . That is, for the current round, we collect quality scores q_{ij} of all pairs $\langle w_i, t_j \rangle$ as samples, and use them to represent probabilistic distributions of quality scores of both worker \widehat{w}_i and task \widehat{t}_j in the future round.

Existence Probabilities of Pairs with the Predicted Workers/Tasks. Note that, some assignment pairs that involve the predicted worker/task may not be valid, due to the time constraints of spatial tasks t_j or \widehat{t}_j (i.e., deadline e_j). Thus, we will associate each pair (with either worker or task in future) with an existence probability, \widehat{p}_{ij} .

For pair $\langle \widehat{w}_i, t_j \rangle$, we let $\widehat{p}_{ij} = \min\{\frac{n_i}{|W_p|}, 1\}$, where n_i is the number of valid workers who can reach task t_j at the current timestamp p , and $|W_p|$ is the total number of (estimated) workers at timestamp p .

Similarly, for pair $\langle w_i, \hat{t}_j \rangle$, we can obtain: $\hat{p}_{ij} = \min\{\frac{m_j}{|T_p|}, 1\}$, where m_j is the number of valid tasks that worker w_i can reach before the deadlines, and $|T_p|$ is the total number of (estimated) tasks at timestamp p .

For pair $\langle \hat{w}_i, \hat{t}_j \rangle$, let u_{ij} be the total number of valid pairs between W_p and T_p at the current timestamp p . Then, we can estimate the existence probability of pair $\langle \hat{w}_i, \hat{t}_j \rangle$ by: $\hat{p}_{ij} = \frac{u_{ij}}{|W_p| \cdot |T_p|}$.

IV. THE GREEDY APPROACH

In this section, we propose an efficient greedy algorithm to solve the PB-SC problem, which iteratively finds one “best” worker-and-task assignment pair, $\langle w_i, t_j \rangle$, each time, with the highest increase of the quality score and under the budget constraint of high confidences. Here, in order to guarantee the global optimality, our greedy algorithm is applied over both current and predicted future workers/tasks (i.e., in the current and next rounds at timestamps p and $(p+1)$, respectively).

After all assignment pairs (involving current/future workers/tasks) are selected, we will only insert into set I_p those pairs, $\langle w_i, t_j \rangle$, with both workers and tasks at current timestamp p (i.e., $w_i \in W_p$ and $t_j \in T_p$), which is expected to achieve the global optimality of assignments. This process exactly corresponds to line 6 of procedure PB-SC_Framework in Figure 3.

A. The Comparisons of the Quality Score Increases / Traveling Cost Increases

Since our greedy algorithm needs to select one worker-and-task assignment pair, $\langle \tilde{w}_i, \tilde{t}_j \rangle$, at a time with the highest increase of the total quality score, in this subsection, we will first formalize the increase of the quality score, $\Delta_q(\tilde{w}_i, \tilde{t}_j)$, for a pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$, and then compare the increases of total quality scores between two pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ and $\langle \tilde{w}_a, \tilde{t}_b \rangle$, where $\tilde{w}_i, \tilde{t}_j, \tilde{w}_a$, and \tilde{t}_b can be either current or predicted workers/tasks.

The Calculation of the Quality Score Increase, $\Delta_q(\tilde{w}_i, \tilde{t}_j)$. Based on Eq. (1), the total quality score is given by summing up all quality scores of the selected assignment pairs. Thus, when we choose a new assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$, the increase of the quality score, $\Delta_q(\tilde{w}_i, \tilde{t}_j)$, is exactly equal to the quality score of this new pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$, denoted as \tilde{q}_{ij} . That is,

$$\Delta_q(\tilde{w}_i, \tilde{t}_j) = \tilde{q}_{ij}, \quad (6)$$

where \tilde{q}_{ij} is a fixed value, if both \tilde{w}_i and \tilde{t}_j are current worker and task, respectively; otherwise, \tilde{q}_{ij} is a random variable whose distribution can be given by samples discussed in Section III-B.

The Comparisons of the Quality Score Increase Between Two Pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ and $\langle \tilde{w}_a, \tilde{t}_b \rangle$. Next, we discuss how to decide which worker-and-task assignment pair is better, in terms of the quality score increase, between two pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ and $\langle \tilde{w}_a, \tilde{t}_b \rangle$.

Specifically, if both pairs have workers/tasks at current timestamp p , then the quality score increases, \tilde{q}_{ij} and \tilde{q}_{ab} (given in Eq. (6)), are fixed values. In this case, the pair with higher quality score increase is better.

On the other hand, in the case that either of the two pairs involves the predicted workers/tasks, their corresponding quality score increases, that is, \tilde{q}_{ij} and/or \tilde{q}_{ab} , are random variables. To compare the two quality score increases, we can compute the probability, $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}$, that pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ has the

increase greater than that of the other one. That is, by applying the *central limit theorem* (CLT) [14], [16], we have:

$$\begin{aligned} Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)} &= Pr\{\Delta_q(\tilde{w}_i, \tilde{t}_j) > \Delta_q(\tilde{w}_a, \tilde{t}_b)\} \\ &= Pr\{\tilde{q}_{ij} > \tilde{q}_{ab}\} \\ &= 1 - Pr\{\tilde{q}_{ij} \leq \tilde{q}_{ab}\} \\ &= 1 - Pr\left\{\frac{\tilde{q}_{ij} - \tilde{q}_{ab} - (E(\tilde{q}_{ij}) - E(\tilde{q}_{ab}))}{\sqrt{Var(\tilde{q}_{ij}) + Var(\tilde{q}_{ab})}} \leq \frac{-(E(\tilde{q}_{ij}) - E(\tilde{q}_{ab}))}{\sqrt{Var(\tilde{q}_{ij}) + Var(\tilde{q}_{ab})}}\right\} \\ &= 1 - \Phi\left(\frac{-(E(\tilde{q}_{ij}) - E(\tilde{q}_{ab}))}{\sqrt{Var(\tilde{q}_{ij}) + Var(\tilde{q}_{ab})}}\right), \end{aligned} \quad (7)$$

where $\Phi(\cdot)$ is the *cumulative density function* (cdf) of a standard normal distribution.

By using Eq. (7), we can compute the probability, $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}$, that pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ is better than (i.e., with higher score than) pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$. If it holds that $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)} > 0.5$, then we say that pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ is expected to have higher quality score increase; otherwise, pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$ has higher quality score increase.

The Comparisons of the Traveling Cost Increase Between Two Pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ and $\langle \tilde{w}_a, \tilde{t}_b \rangle$. Similar to the quality score, we can also compute the probability, $Pr_{\Delta_c(\tilde{w}_i, \tilde{t}_j)}$, that the increase of the traveling cost for pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ is smaller than that of pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$. That is, we can obtain:

$$\begin{aligned} Pr_{\Delta_c(\tilde{w}_i, \tilde{t}_j)} &= Pr\{\Delta_c(\tilde{w}_i, \tilde{t}_j) \leq \Delta_c(\tilde{w}_a, \tilde{t}_b)\} \\ &= Pr\{\tilde{c}_{ij} \leq \tilde{c}_{ab}\} \\ &= \Phi\left(\frac{-(E(\tilde{c}_{ij}) - E(\tilde{c}_{ab}))}{\sqrt{Var(\tilde{c}_{ij}) + Var(\tilde{c}_{ab})}}\right). \end{aligned} \quad (8)$$

B. The Pruning Strategy

As discussed in Section IV-A, one straightforward method for selecting a “good” assignment pair at a time is as follows. We sequentially scan each valid worker-and-task pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$, and compare its quality score increase, $\Delta_q(\tilde{w}_i, \tilde{t}_j)$, with that of the best-so-far pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$, in terms of the probability $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}$. If the pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ expects to have higher quality score (and moreover satisfy the budget constraint), then we consider it as the new best-so-far pair.

The straightforward method mentioned above considers all possible valid assignment pairs, and computes their comparison probabilities, which requires high time complexity, that is, $O(m' \cdot n')$, where m' and n' are the numbers of tasks and workers in both current and future rounds, respectively. Therefore, in this subsection, we will propose effective pruning methods to quickly discard those false alarms of assignment pairs, with both high traveling costs and low quality scores.

Pruning with Bounds of Quality and Traveling Cost.

Without loss of generality, for each pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$, assume that we can obtain its lower and upper bounds of the traveling cost, \tilde{c}_{ij} , and quality score, \tilde{q}_{ij} , where \tilde{c}_{ij} and \tilde{q}_{ij} are either fixed values (if \tilde{w}_i and \tilde{t}_j are worker/task at the current round) or random variables (if worker and/or task are from future round). That is, we denote $\tilde{c}_{ij} \in [lb_{\tilde{c}_{ij}}, ub_{\tilde{c}_{ij}}]$ and $\tilde{q}_{ij} \in [lb_{\tilde{q}_{ij}}, ub_{\tilde{q}_{ij}}]$.

This way, in a 2D quality-and-travel-cost space, each worker-and-task assignment pair, $\langle \tilde{w}_i, \tilde{t}_j \rangle$, corresponds to a rectangle, $[lb_{\tilde{q}_{ij}}, ub_{\tilde{q}_{ij}}] \times [lb_{\tilde{c}_{ij}}, ub_{\tilde{c}_{ij}}]$. Then, based on the idea of the *skyline* query [6], we can safely prune those pairs that are *dominated* by candidate pairs, in terms of the traveling cost and quality score.

Lemma 4.1: (The Dominance Pruning) Given a candidate pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$, a valid worker-and-task pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ can be

safely pruned, if and only if it holds that: (1) $ub_c_{ab} < lb_c_{ij}$, and (2) $lb_q_{ab} > ub_q_{ij}$.

Proof: Since it holds that $c_{ij} \in [lb_c_{ij}, ub_c_{ij}]$ and $q_{ij} \in [lb_q_{ij}, ub_q_{ij}]$, by lemma assumptions and inequality transition, we have:

$$c_{ab} \leq ub_c_{ab} < lb_c_{ij} \leq c_{ij}, \text{ and}$$

$$q_{ab} \geq lb_q_{ab} > ub_q_{ij} \geq q_{ij}.$$

As a result, we can see that, compared to pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$, pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ has both higher traveling cost c_{ij} and lower quality score q_{ij} . Since our greedy algorithm only selects one best pair each time (which can maximally increase the quality score and minimally increase the traveling cost), pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ is definitely worse than $\langle \tilde{w}_a, \tilde{t}_b \rangle$ in both quality and traveling cost dimensions, and thus can be safely pruned. ■

Pruning with the Increase Probability. Lemma 4.1 utilizes the lower/upper bounds of the quality score and traveling cost to enable the dominance pruning. If a pair cannot be simply pruned by Lemma 4.1, we will further consider a more costly pruning method, by consider the probabilistic information.

Lemma 4.2: (The Increase Probability Pruning) Given a candidate pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$, a valid pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ can be safely pruned, if and only if it holds that: (1) $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}$ (w.r.t. $\langle \tilde{w}_a, \tilde{t}_b \rangle$) is greater than 0.5, and (2) $Pr_{\Delta_c(\tilde{w}_i, \tilde{t}_j)}$ (w.r.t. candidate pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$) is greater than 0.5, where $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}$ and $Pr_{\Delta_c(\tilde{w}_i, \tilde{t}_j)}$ are given by Eqs. (7) and (8), respectively.

Intuitively, Lemma 4.2 filters out those pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ that have higher probabilities to be inferior to other candidate pairs $\langle \tilde{w}_a, \tilde{t}_b \rangle$, in terms of both traveling cost and quality score.

Based on Lemmas 4.1 and 4.2, we can obtain a set, S_p , of candidate pairs that cannot be dominated by other pairs.

Selection of the Best Pair Among Candidate Pairs. Given a number of candidate pairs in set S_p , our greedy algorithm still needs to identify one “best” pair with high quality score and under the budget constraint. Specifically, we will first filter out those false alarms in S_p with high traveling cost (i.e., violating the budget constraint), and then return one pair with the highest probability to have larger quality score than others in S_p .

Assume that in the greedy algorithm, we have so far selected L pairs, denoted as $\langle \tilde{w}_a, \tilde{t}_b \rangle$. Then, with a new assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle \in S_p$, if it holds that:

$$Pr \left\{ \left(\sum_{\forall \langle \tilde{w}_a, \tilde{t}_b \rangle} lb_c_{ab} \right) + c_{ij} \leq B_{max} \right\} \leq \delta, \quad (9)$$

then pair $\langle \tilde{w}_i, \tilde{t}_j \rangle \in S_p$ can be safely ruled out from candidate set S_p , where δ is a user-specified confidence level that the selected assignment satisfies the budget constraint B_{max} for both (remaining) current- and next-round budgets. Eq. (9) can be computed via CLT [14], [16].

Next, in the remaining candidate pairs in S_p , we will select one pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ with the highest probability, $Pr_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle)$, of having the largest high quality. That is, we have:

$$Pr_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle) = \prod_{\forall \langle \tilde{w}_a, \tilde{t}_b \rangle} Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}(\langle \tilde{w}_a, \tilde{t}_b \rangle) \quad (10)$$

where $Pr_{\Delta_q(\tilde{w}_i, \tilde{t}_j)}(\langle \tilde{w}_a, \tilde{t}_b \rangle)$ is the probability of quality score increase, compared with pair $\langle \tilde{w}_a, \tilde{t}_b \rangle$, given by Eq. (7).

Finally, among all the remaining candidate pairs in set S_p , we will choose the one, $\langle \tilde{w}_i, \tilde{t}_j \rangle$, with the highest probability

Procedure PB-SC_Greedy {

Input: current and predicted workers \tilde{w}_i in W , current and predicted tasks \tilde{t}_j in T , and the maximum possible budget B_{max}
Output: a worker-and-task assignment instance set, I_p
(1) $I_p = \emptyset$;
(2) obtain a list, \mathcal{L} , of valid worker-and-task pairs for $\tilde{w}_i \in W$ and $\tilde{t}_j \in T$
(3) **for** $k = 1$ to $\min\{|W|, |T|\}$
(4) $S_p = \emptyset$;
(5) **for** each valid assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle \in \mathcal{L}$
(6) **if** $\langle \tilde{w}_i, \tilde{t}_j \rangle$ has lb_c_{ij} greater than the remaining budget, then **continue**;
(7) **if** pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ cannot be pruned w.r.t. S_p by Lemma 4.1
(8) **if** pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ cannot be pruned w.r.t. S_p by Lemma 4.2
(9) add $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to S_p
(10) prune other candidate pairs in S_p with $\langle \tilde{w}_i, \tilde{t}_j \rangle$
(11) select one best assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ in S_p satisfying the budget constraint B_{max} in Eq. (9) and with the highest probability $Pr_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle)$ in Eq. (10)
(12) add the selected pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to I_p
(13) remove all those valid pairs $\langle \tilde{w}_i, - \rangle$ or $\langle -, \tilde{t}_j \rangle$ from \mathcal{L}
(14) remove those worker-and-task pairs with the predicted workers/tasks from I_p
(15) **return** I_p }

Fig. 4. The Greedy Algorithm.

$Pr_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle)$, which will be included as a selected best assignment pair in our greedy algorithm.

C. The Greedy Algorithm

In this subsection, we propose a greedy algorithm, which iteratively assigns a worker to a spatial task that can maximize the total quality score under the budget constraint each time.

Figure 4 presents the pseudo code of our PB-SC greedy algorithm, namely PB-SC_Greedy, which obtains one best worker-and-task assignment pair each time over both current and predicted future workers and tasks, where the selected pair satisfies the budget constraint B_{max} and has the largest quality score with high confidence, where B_{max} is the available budget in both current and next rounds.

We first initialize the worker-and-task assignment instance set I_p with an empty set (line 1). Then, we obtain a list, \mathcal{L} , of valid worker-and-task assignment pairs, which may involve either current or future workers/tasks, that is, $\tilde{w}_i \in W$ and $\tilde{t}_j \in T$ (line 2). Next, for each iteration, we find one best assignment pair with high quality score and low traveling cost (satisfying the budget constraint) (lines 3-13). In particular, we check each valid assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ in the list \mathcal{L} (line 5). If this pair has the lower bound, lb_c_{ij} , of the traveling cost greater than (the upper bound of) the remaining budget (w.r.t. I_p and B_{max}), then it does not satisfy the budget constraint, and we can continue to check the next assignment pair (line 6). Then, if the pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ cannot be pruned by dominance and increase probability pruning methods in Lemmas 4.1 and 4.2, respectively, then $\langle \tilde{w}_i, \tilde{t}_j \rangle$ is a candidate pair, and we include in an initially empty candidate set S_p (lines 7-9). In addition, we can also use candidate pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to prune other pairs in set S_p (line 10). After that, we can insert the best pair from the candidate set S_p into set I_p (lines 11-12), such that the budget constraint B_{max} is satisfied in Eq. (9) and the probability $Pr_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle)$ in Eq. (10) is maximized. Since each worker can be assigned with at most one task and each task is accomplished by at most one worker, we remove those valid pairs from \mathcal{L} that contains either worker \tilde{w}_i or task \tilde{t}_j (line 13). Finally, we remove those worker-and-task pairs involving future workers/tasks from I_p , and return the set I_p as the solution of the greedy algorithm (lines 14-15).

V. THE DIVIDE-AND-CONQUER APPROACH

In this section, we propose an efficient *divide-and-conquer algorithm* (D&C), which partitions the PB-SC problem into g subproblems, recursively conquers the subproblems, and merges assignment results from subproblems. In this paper,

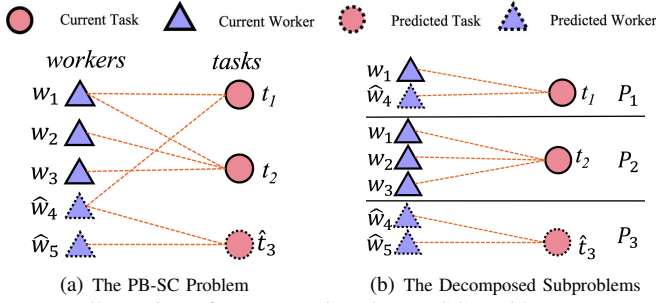


Fig. 5. Illustration of Decomposing the PB-SC Problem. we will divide the PB-SC problem with m' current/future tasks into g subproblems, each involving $\lceil m'/g \rceil$ tasks. The D&C process continues, until the subproblem sizes become 1 (i.e., with one single spatial task in subproblems, which can be easily solved by our greedy algorithm). We will later discuss how to utilize a cost model to estimate the best g value that can achieve low PB-SC processing cost.

A. The Decomposition of the PB-SC Problem

In this subsection, we discuss how to decompose a PB-SC problem into subproblems, with respect to tasks.

Decomposing the PB-SC Problem. Specifically, assume that the original PB-SC problem involves m' current/future spatial tasks for both current and next rounds. Our goal is to divide this problem into g subproblems M_s (for $1 \leq s \leq g$), such that each subproblem M_s involves a disjoint subgroup of $\lceil m'/g \rceil$ spatial tasks, \tilde{t}_j , each of which is associated with potentially valid worker(s) \tilde{w}_i (i.e., with valid worker-and-task assignment pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$).

Note that, after the decomposition, each subproblem M_s contains all valid pairs, $\langle \tilde{w}_i, \tilde{t}_j \rangle$, w.r.t. the decomposed $\lceil m'/g \rceil$ tasks. Since tasks in different subproblems may be reachable by the same workers, different subproblems may involve the same (conflicting) workers, whose conflicts should be resolved when we merge solutions (the selected assignment pairs) to these subproblems (as will be discussed in Section V-B).

Example 3 (The PB-SC Problem Decomposition) Figure 5 shows an example of decomposing the PB-SC problem (as shown in Figure 5(a)) into 3 subproblems (as depicted in Figure 5(b)), where each subproblem contains one single spatial task (i.e., subproblem size = 1), associated with its related valid workers. Here, the dashed border indicates the predicted future workers (i.e., w_4 and w_5) or task (i.e., t_3). In this example, the first subproblem in Figure 5(b) contains task t_1 , which can be reached by workers w_1 and w_4 . Different tasks may have conflicting workers, for example, tasks t_1 and t_2 from subproblems M_1 and M_2 , respectively, share the same (conflicting) worker w_1 .

The PB-SC Decomposition Algorithm. Figure 6 illustrates the pseudo code of our PB-SC decomposition algorithm, namely PB-SC_Decomposition, which decomposes the PB-SC problem (with m' tasks), and returns g PB-SC subproblems, M_s (each having $\lceil m'/g \rceil$ tasks).

Specifically, we first initialize g empty subproblems, M_s , where $1 \leq s \leq g$ (lines 1-2). Then, we find all valid worker-and-task assignment pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ for both current and predicted workers/tasks in sets W and T , respectively (line 3).

Next, we want to iteratively retrieve g subproblems, M_s , from the original PB-SC problem (lines 4-8). That is, for the s -th iteration, we first obtain an anchor task \tilde{t}_j and its $(\lceil m'/g \rceil - 1)$ nearest tasks, and add them to set $T_p^{(s)}$ (line 5),

Procedure PB-SC_Decomposition {

Input: m' current/future workers \tilde{w}_i in W , m' current/future spatial tasks \tilde{t}_j in T , and the number of subproblems g
Output: the decomposed PB-SC subproblems, M_s (for $1 \leq s \leq g$)
(1) **for** $s = 1$ to g
(2) $M_s = \emptyset$
(3) compute all valid worker-and-task pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ from W and T
(4) **for** $s = 1$ to g
(5) add an anchor task \tilde{t}_j and find its $(\lceil m'/g \rceil - 1)$ nearest tasks to set $T_p^{(s)}$ // the task, \tilde{t}_j , whose longitude (or mean of the longitude) is the smallest
(6) **for** each current/future task $\tilde{t}_j \in T_p^{(s)}$
(7) obtain all valid workers \tilde{w}_i that can reach task \tilde{t}_j
(8) add these pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to subproblem M_s
(9) **return** subproblems M_1, M_2, \dots , and M_g }

Fig. 6. The PB-SC Problem Decomposition Algorithm.

where anchor tasks \tilde{t}_j are chosen in a *sweeping style* (starting with the smallest longitude, or mean of the longitude for future tasks; in the case that multiple tasks have the same longitude, we choose the one with smallest latitude).

For each task $\tilde{t}_j \in T_p^{(s)}$, we obtain all its related workers \tilde{w}_i who can reach task \tilde{t}_j , and add pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to subproblem M_s (lines 6-8). Finally, we return the g decomposed subproblems M_s (for $1 \leq s \leq g$) (line 9).

B. The PB-SC Merge Algorithm

As mentioned earlier, we can execute the decomposition algorithm to recursive divide the PB-SC problem, until each subproblem only involves one single task (which can be easily processed by the greedy algorithm). After we obtain solutions to the decomposed PB-SC subproblems (i.e., a number of selected assignment pairs in subproblems), we need to merge these solutions into the one to the original PB-SC problem.

Resolving Worker-and-Task Assignment Conflicts. During the merge process, some workers are conflicting, that is, they are assigned to different tasks in the solutions to distinct subproblems at the same time. This contradicts with the requirement that each worker can only be assigned with at most one spatial task at a time. Thus, in order to merge solutions to these (conflicting) subproblems, we need to resolve the conflicts.

Next, we use an example to illustrate how to resolve the conflicts between two (or multiple) pairs $\langle \tilde{w}_i, \tilde{t}_j \rangle$ and $\langle \tilde{w}_i, \tilde{t}_b \rangle$ (w.r.t. the conflicting worker \tilde{w}_i), by selecting one “best” pair with low traveling cost and high quality score.

Example 4 (The Merge of Subproblems) In the example of Figure 5(b), assume that in subproblems M_1 and M_2 , we selected pairs $\langle w_1, t_1 \rangle$ and $\langle w_1, t_2 \rangle$ as the best assignment, respectively, which contains the conflicting worker w_1 . When we merge the two subproblems M_1 and M_2 , we need to resolve such a conflict by deciding which task should be assigned to the conflicting worker w_1 . By using Lemmas 4.1 and 4.2, we can first prune pairs that are not dominated by others. Then, among the remaining candidate pairs, we can select one best pair satisfying Eq. (9) and maximizing Eq. (10). In this example, assume that pair $\langle w_1, t_1 \rangle$ dominates pair $\langle w_1, t_2 \rangle$ by Lemma 4.1. Then, we will assign worker w_1 with task t_1 (since $\langle w_1, t_1 \rangle$ is the best pair), and find another worker (e.g., w_2) with the largest quality score under the budget constraint to do task t_2 . This way, after solving conflicts, we obtain two updated pairs $\langle w_1, t_1 \rangle$ and $\langle w_2, t_2 \rangle$ for subproblems M_1 and M_2 , respectively.

The PB-SC Merge Algorithm. Figure 7 illustrates the PB-SC merge algorithm, namely PB-SC_Merge, which resolves the conflicts between the current assignment instance set I_p (that we have merged subproblems M) and that, $I_p^{(s)}$, of subproblem M_s , and returns a merged set without conflicts.

Procedure PB-SC_Merge {
Input: the current assignment instance set, I_p , of the merged subproblems M , and the assignment instance set, $I_p^{(s)}$, of subproblem M_s
Output: a merged worker-and-task assignment instance set, I_p
(1) let W_c be a set of conflicting workers between I_p and $I_p^{(s)}$
(2) **while** $W_c \neq \emptyset$
(3) choose a worker $\tilde{w}_i \in W_c$ with the highest traveling cost in $I_p^{(s)}$
// assume \tilde{w}_i is assigned to \tilde{t}_j in I_p and to \tilde{t}_k in $I_p^{(s)}$
(4) select one best pair between $\langle \tilde{w}_i, \tilde{t}_j \rangle \in I_p$ and $\langle \tilde{w}_i, \tilde{t}_k \rangle \in I_p^{(s)}$
// by using Lemmas 4.1 and 4.2, and finding the one satisfying
// Eq. (9) and maximizing Eq. (10)
(5) **if** pair $\langle \tilde{w}_i, \tilde{t}_k \rangle$ in subproblem M_s is selected
(6) find another best worker \tilde{w}'_i in M and substitute $\langle \tilde{w}'_i, \tilde{t}_j \rangle$ in I_p
(7) **else**
(8) find another best worker \tilde{w}''_i in M_s and substitute $\langle \tilde{w}''_i, \tilde{t}_k \rangle$ in $I_p^{(s)}$
(9) $W_c = W_c - \{\tilde{w}_i\}$
(10) $I_p = I_p \cup I_p^{(s)}$
(11) **return** I_p }

Fig. 7. The Merge Algorithm.

First, we obtain a set, W_c , of conflicting workers between I_p and $I_p^{(s)}$ (line 1), which are assigned with different tasks in different subproblems, M and M_s . Then, in each iteration, we select one conflicting worker $\tilde{w}_i \in W_c$ with the highest traveling cost in $I_p^{(s)}$, and choose one best pair between $\langle \tilde{w}_i, \tilde{t}_j \rangle \in I_p$ and $\langle \tilde{w}_i, \tilde{t}_k \rangle \in I_p^{(s)}$, in terms of budget and quality score (which can be achieved by checking Lemmas 4.1 and 4.2, and finding the one that satisfies Eq. (9) and maximizes Eq. (10) (lines 2-4).

When $\langle \tilde{w}_i, \tilde{t}_k \rangle$ in subproblem M_s is selected as the best pair, we can resolve the conflicts by replacing $\langle \tilde{w}_i, \tilde{t}_j \rangle$ with $\langle \tilde{w}'_i, \tilde{t}_j \rangle$ in I_p ; otherwise, we can replace $\langle \tilde{w}_i, \tilde{t}_k \rangle$ with $\langle \tilde{w}''_i, \tilde{t}_k \rangle$ in $I_p^{(s)}$ (lines 5-8). Then, we remove worker \tilde{w}_i from set W_c (line 9).

After resolving all conflicting workers in W_c between I_p and $I_p^{(s)}$, we can union them together, and return an updated merged set I_p (lines 10-11).

C. The D&C Algorithm

Up to now, we have discussed how to decompose and merge subproblems. In this subsection, we will illustrate the detailed *divide-and-conquer* (D&C) algorithm, which partitions the original PB-SC problem into subproblems, recursively solves each subproblem, and merges assignment results of subproblems by resolving conflicts and adjusting the assignments under the budget constraint.

Figure 8 shows the pseudo code of our D&C algorithm, namely procedure PB-SC_D&C. We first initialize an empty set r_{lt} , which is used for store candidate pairs chosen by our D&C algorithm (line 1). Then, we utilize a novel cost model (discussed in Appendix C to estimate the best number of the decomposed subproblems, g , with respect to current/future sets, W and T , of workers and tasks, respectively (line 2). With this parameter g , we can invoke the PB-SC_Decomposition algorithm (as mentioned in Figure 6), and obtain g subproblems M_s (line 3).

For each subproblem M_s , if M_s involves more than 1 task, then we can recursively call procedure PB-SC_D&C (\cdot) to obtain the best assignment pairs from subproblem M_s (lines 4-6). Otherwise, if subproblem M_s only contains one single spatial task \tilde{t}_j , then we apply the greedy algorithm (in Figure 4) to select one “best” worker for task \tilde{t}_j (lines 7-8). Here, the best worker means, the corresponding pair has the highest quality score under the budget constraint B_{max} .

After that, the selected assignment pairs in the s -th subproblem are kept in set $r_{lt}^{(s)}$, where $1 \leq s \leq g$. Then,

Procedure PB-SC_D&C {
Input: n' current/future workers in W , and m' current/future spatial tasks in T , and a maximum budget B_{max}
Output: an assignment instance set, I_p , with current/future workers/tasks
(1) $r_{lt} = \emptyset$
(2) estimate the best number of subproblems, g , w.r.t. W and T
(3) invoke PB-SC_Decomposition(W, T, g), and obtain g subproblems M_s
(4) **for** $s = 1$ to g
(5) **if** the number of tasks in subproblem M_s is greater than 1
(6) $r_{lt}^{(s)} = \text{PB-SC_D\&C}(W(M_s), T(M_s), B_{max})$
(7) **else**
(8) $r_{lt}^{(s)} = \text{PB-SC_Greedy}(W(M_s), T(M_s), B_{max})$
(9) **for** $s = 1$ to g
(10) find the next subproblem, M_s
(11) $r_{lt} = \text{PB-SC_Merge}(r_{lt}, r_{lt}^{(s)})$
(12) **if** the upper bound of the traveling cost of $r_{lt} \leq B_{max}$
(13) **return** r_{lt}
(14) **else**
(15) $I_p = \text{PB-SC_Budget_Constrained_Selection}(r_{lt}, B_{max})$
(16) **return** I_p }

Procedure PB-SC_Budget_Constrained_Selection {
Input: candidate pairs in r_{lt} and a maximum budget B_{max}
Output: a worker-and-task assignment instance set, I_p , under the budget constraint
(17) $I_p = \emptyset$;
(18) **for** $k = 1$ to $|r_{lt}|$
(19) $S_p = \emptyset$
(20) **for** each assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle \in r_{lt}$
(21) **if** $\langle \tilde{w}_i, \tilde{t}_j \rangle$ has $lb_{c_{ij}}$ greater than the remaining budget, then continue;
(22) **if** pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ cannot be pruned w.r.t. S_p by Lemma 4.1
(23) **if** pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ cannot be pruned w.r.t. S_p by Lemma 4.2
(24) add $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to S_p
(25) prune other candidate pairs in S_p with $\langle \tilde{w}_i, \tilde{t}_j \rangle$
(26) select one best assignment pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ in S_p satisfying the budget constraint B_{max} in Eq. (9) and with the highest probability $M_{q,max}(\langle \tilde{w}_i, \tilde{t}_j \rangle)$ in Eq. (10)
(27) add the selected pair $\langle \tilde{w}_i, \tilde{t}_j \rangle$ to I_p
(28) **return** I_p }

Fig. 8. The Divide-and-Conquer Algorithm.
we can invoke procedure PB-SC_Merge (\cdot) to merge these g sets $r_{lt}^{(s)}$ into a set r_{lt} , by resolving the conflicts (lines 9-11).

Due to the budget constraint B_{max} , we may still need to adjust assignment pairs in set r_{lt} such that the total traveling cost is below the maximum budget B_{max} . If the upper bound of the traveling cost in set r_{lt} does not exceed budget B_{max} , then we can directly return r_{lt} as I_p (lines 12-13). Otherwise, similar to the greedy algorithm, we need to select “best” assignment pairs from set r_{lt} that are under the budget constraint B_{max} (maximizing the total quality score), and add them to the set I_p , by calling procedure PB-SC_Budget_Constrained_Selection (lines 14-15).

In particular, to adjust the budget, we select a “best” pair from set r_{lt} each time that satisfies the budget constraint B_{max} and with the highest quality score, similar to the greedy algorithm. Please refer to details of procedure PB-SC_Budget_Constrained_Selection in lines 17-28 of Figure 8.

Discussions on Estimating the Best Number, g , of the Decomposed Subproblems. In order to reduce the computation cost in our D&C algorithm, we aim to select a best g value that minimizes the processing cost, in light of our proposed cost model. Specifically, we will formally model the computation cost, $cost_{D\&C}$, of the D&C algorithm, with respect to g , take the derivative of $cost_{D\&C}$ over g , and then let the derivative be 0. This way, we can find the best g value that minimizes the cost of the D&C algorithm. For details, please refer to Appendix B of our technical report [9].

VI. EXPERIMENTAL STUDY

Real/Synthetic Data Sets. We tested our proposed PB-SC processing algorithms over both real and synthetic data sets. Specifically, for real data sets, we used two check-in data sets, Gowalla [11] and Foursquare [21]. In the Gowalla data

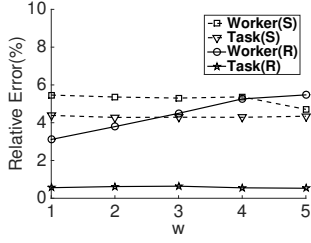


Fig. 9. The Prediction Accuracy vs. Window Size w (Real/Synthetic Data).

set, there are 196,591 nodes (users), with 6,442,890 check-in records. In the Foursquare data set, there are 2,153,471 users, 1,143,092 venues, and 1,021,970 check-ins, extracted from the Foursquare application through the public API. Since most assignments happen in the same cities, we extract check-in records within the area of San Francisco (with latitude from 37.709° to -122.503° and longitude from 37.839° to -122.373°), which has 8,481 Foursquare check-ins, and 149,683 Gowalla check-ins for 6,143 users. We use check-in records of Foursquare to initialize the location and arrival time of tasks in the spatial crowdsourcing system, and configure workers using the check-ins records from Gowalla. In other words, we have 6,143 workers and 8,481 spatial tasks in the experiments over real data. For simplicity, we first linearly map check-in locations from Gowalla and Foursquare into a $[0, 1]^2$ data space, and then scale the arrival times of workers/tasks in the real data accordingly. In order to generate workers/tasks for each round, we evenly divide the entire time span of check-ins from two real data sets into R subintervals ($\in P$), and utilize check-ins in each subinterval to initialize workers/tasks for the corresponding round.

For synthetic data sets, we generate workers/tasks that join the spatial crowdsourcing system for every round within a time interval P as follows. We randomly produce locations of workers and tasks in a 2D data space $[0, 1]^2$, following Gaussian $\mathcal{N}(0.5, 1^2)$ and Zipf distributions (skewness = 0.3), respectively. We also test synthetic worker/task data with other distribution combinations (e.g., Uniform-Zipf) with similar results (see Appendix D of our technical report [9]).

For both real and synthetic data sets, we simulate the velocity v_i of each worker w_i with Gaussian distribution $\mathcal{N}(\frac{v^- + v^+}{2}, (v^+ - v^-)^2)$ within range $[v^-, v^+]$, for $0 < v^- \leq v^+ < 1$, and the unit price C w.r.t. the traveling distance $dist(\cdot, \cdot)$ varies from 5 to 25. Regarding the time constraint (i.e., the deadline e_j) of spatial tasks t_j , we produce the arrival deadlines of tasks within the range $[e^-, e^+]$, which are given by the remaining time for workers to arrive at tasks after these tasks join the system. Moreover, for the total quality score, q_{ij} , of worker-and-task assignments, we randomly generate q_{ij} with Gaussian distributions within $[q^-, q^+]$. In our experiments, we also test the size, w , of the sliding window with values from 1 to 5, and the number, R , of rounds in the time interval P from 10 to 25.

Measures and Competitors. We evaluate the effectiveness and efficiency of our PB-SC processing approaches, in terms of the total quality score and the CPU time. Specifically, the total quality score is defined in Eq. (1), which can measure the quality of the assignment strategy, and the CPU time is given by the average time cost of performing PB-SC assignments for each round.

In our PB-SC problem, regarding the effectiveness, we will compare our PB-SC approach with a straightforward method

TABLE III. EXPERIMENTAL SETTINGS.

Parameters	Values
the size of sliding windows w	1, 2, 3, 4, 5
the budget B	100, 200 , 300, 400, 500
the quality range $[q^-, q^+]$	[0.25, 0.5], [0.5, 1], [1 , 2], [2, 3], [3, 4]
the deadline range $[e^-, e^+]$	[0.25, 0.5], [0.5, 1], [1 , 2], [2, 3], [3, 4]
the velocity range $[v^-, v^+]$	[0.1, 0.2], [0.2 , 0.3], [0.3, 0.4], [0.4, 0.5]
the unit price w.r.t. distance C	5, 10 , 15, 20
the number, R , of rounds	10, 15 , 20, 25
the number, m , of tasks	1K, 3K , 5K, 8K, 10K
the number, n , of workers	1K, 3K , 5K, 8K, 10K

that conducts the assignment over current and future rounds separately (i.e., without prediction), in terms of the total quality score. Moreover, we will also compare our PB-SC approaches (i.e., the greedy and D&C algorithms) with a random method (which randomly assigns workers to spatial tasks under the budget constraint). Since the random method does not take into account the quality of tasks, it is expected to achieve worse quality than our PB-SC approaches (although it is expected to be more efficient than PB-SC).

Experimental Settings. Table III shows our experimental settings, where the default values of parameters are in bold font. In subsequent experiments, each time we vary one parameter, while setting others to their default values. All our experiments were run on an Intel Xeon X5675 CPU @3.07 GHZ with 32 GB RAM in Java.

A. Effectiveness of the PB-SC Approach

The Comparison of the Prediction Accuracy. We first evaluate the prediction accuracy of future workers/tasks in our PB-SC approach, by comparing the estimated numbers, est , of workers/tasks in cells with actual ones, act , in terms of the relative error (i.e., defined as $\frac{|est - act|}{act}$), where the size, w , of the sliding window to do the prediction (via linear regressions) varies from 1 to 5. As shown in Figure 9, for both synthetic (marked with S) and real data (marked with R), relative errors for different window sizes w are not very sensitive to w . Only for real data, the relative error of predicting the number of workers slightly increases with larger w value. This is because the distribution of workers changes quickly over time in real data, which leads to larger prediction error by using wider window size w . Nonetheless, relative errors of predicting the number of workers/tasks remain low (i.e., less than 5.5%) over all real/synthetic data for different window sizes w , which indicates good accuracy of our grid-based prediction approach.

Comparison with a Straightforward Method. Figure 10 compares the quality score of our PB-SC approaches (with predicted workers/tasks) with that of the straightforward method which selects assignments in current and next rounds separately (without predictions), where budget B varies from 100 to 500. We denote PB-SC approaches with prediction as GREEDY_WP, D&C_WP, and RANDOM_WP, and those without prediction as GREEDY_WoP, D&C_WoP, and RANDOM_WoP, respectively.

In Figure 10(a), we can see that, the quality scores of PB-SC with predictions (with solid lines) are higher than that of PB-SC without prediction (with dash lines), for different budget B . This indicates the effectiveness of our PB-SC approaches over current/predicted workers/tasks, which can achieve better assignment strategy than the ones without prediction (i.e., local optimality). Moreover, either with or without predictions, D&C incurs higher quality scores than GREEDY (since D&C is carefully designed to find assignments with high quality scores via divide-and-conquer), and RANDOM has the lowest score, which implies good

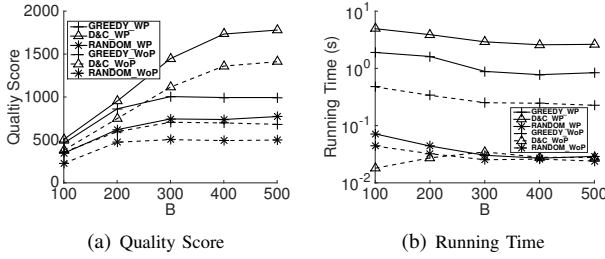


Fig. 10. Effect of the Budget B for Each Round (Synthetic Data).

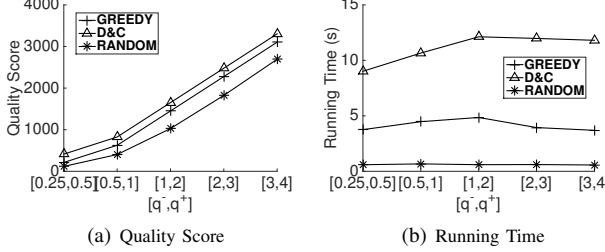


Fig. 11. Effect of the Range of Quality Score q_{ij} (Real Data). quality of our proposed assignment strategies.

Figure 10(b) illustrates the average running time of our GREEDY and D&C approaches, compared with RANDOM, for each round. In the figure, due to the prediction and merge costs, D&C_WP requires the highest CPU time to solve the PB-SC problem, which trades the efficiency for the accuracy. When the budget B increases, the running time of D&C decreases. This is because larger budget B leads to lower cost of selecting assignment pairs under the budget constraint (i.e., lines 12-15 in procedure PB-SC_D&C of Figure 8). For other approaches, the time cost remains low (i.e., less than 5 seconds) for different B values.

B. Performance of the PB-SC Approach

The PB-SC Performance vs. the Range, $[q^-, q^+]$, of Quality Score q_{ij} . Figure 11 illustrates the experimental results on different ranges, $[q^-, q^+]$, of quality score q_{ij} from $[0.25, 0.5]$ to $[3, 4]$ on real data. In Figure 11(a), with the increase of score ranges, quality scores of all the three approaches increase. D&C has higher quality score than GREEDY, which are both higher than RANDOM. From Figure 11(b), RANDOM is the fastest (however, with the lowest quality score), since it randomly selects assignments without considering the quality score maximization. D&C has higher running time than GREEDY (however, higher quality scores than GREEDY). Nonetheless, the running times of GREEDY remain low (i.e., 1-5 seconds).

The PB-SC Performance vs. the Range, $[e^-, e^+]$, of Tasks' Deadlines e_j . Figure 12 shows the effect of the range, $[e^-, e^+]$, of tasks' deadlines e_j on the PB-SC performance over real data, where $[e^-, e^+]$ changes from $[0.25, 0.5]$ to $[2, 3]$. In Figure 12(a), when the range $[e^-, e^+]$ becomes larger, quality scores of all three approaches also increase. Since a more relaxed (larger) deadline e_j of a task t_j can be conducted by more valid workers, it thus leads to higher quality score (that can be achieved) and processing time (as confirmed by Figure 12(b)). Similar to previous results, D&C can achieve higher quality scores than GREEDY, and both of them outperform RANDOM. Furthermore, GREEDY needs higher time cost than RANDOM, and has lower running time than D&C.

The PB-SC Performance vs. the Range, $[v^-, v^+]$, of Workers' Velocities v_i . Figure 13 presents the PB-SC performance with different ranges, $[v^-, v^+]$, of workers' velocities v_i from $[0.1, 0.2]$ to $[0.4, 0.5]$ on synthetic data, where default values are used for other parameters. In

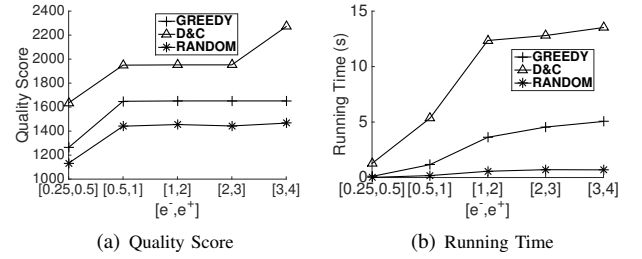


Fig. 12. Effect of the Range of Tasks' Deadlines e_j (Real Data).

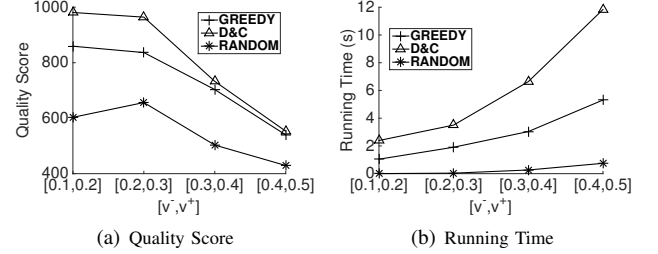


Fig. 13. Effect of the Range of Velocities $[v^-, v^+]$ (Synthetic Data).

Figure 13(a), when the value range, $[v^-, v^+]$, of velocities of workers increases, the total quality scores of all three approaches decrease. When the range of velocities increases, some worker-and-task pairs with long distances may become valid, which may quickly consume the available budget B and in turn reduce the number of selected pairs. Thus, for larger workers' velocities, the resulting total quality score for the selected pairs in GREEDY and D&C approaches decreases. With different velocities, D&C always has higher quality scores than GREEDY, followed by RANDOM. In Figure 13(b), the running times of GREEDY and D&C increase for larger workers' velocities, since there are more worker-and-task assignments to process. Moreover, RANDOM has the smallest time cost (however, the worst quality score), whereas GREEDY runs faster than D&C, and slower than RANDOM.

The PB-SC Performance vs. the Number, m , of Tasks . Figure 14 varies the total number, m , of spatial tasks for R ($= 15$ by default) rounds from $1K$ to $10K$ on synthetic data sets, where other parameters are set to their default values. From the experimental results, with the increase of m , the total quality scores and time costs of all the three approaches both increase smoothly. This is reasonable, since more tasks lead to more valid pairs, which incur higher quality score for selected assignment pairs, and require higher time cost to process. D&C can achieve higher quality score than GREEDY, which indicates the effectiveness of our proposed D&C approach. Moreover, RANDOM has the worst quality score among the three tested approaches.

The PB-SC Performance vs. the Number, n , of Workers. Figure 15 examines the effect of the number, n , of workers for R ($= 15$ by default) rounds on the PB-SC performance, where n varies from $1K$ to $10K$ and other parameters are set to default values. Similarly, both total quality scores and running times of three PB-SC approaches increase, for larger n values. Our GREEDY and D&C algorithms can achieve better quality score than RANDOM. Further, GREEDY has lower time cost than D&C. Nonetheless, the time cost of our approaches smoothly grows with the increasing n values, which indicates good scalability of our PB-SC approaches.

Due to space limitations, please refer to the experimental results on the effects of the Number, R , of Rounds and the Unit Price C w.r.t. Distance $dist(w_i, t_j)$ in Appendix E of the technical report [9].

VII. RELATED WORK

There are many previous works on crowdsourcing systems [5], [7], which usually allow workers to accept task requests and accomplish tasks online. However, these workers do not have to travel to some sites to perform tasks. In contrast, the spatial crowdsourcing system [13], [18] requires workers traveling to locations of spatial tasks, and complete tasks such as taking photos/videos. For instance, some related works [12], [17] studied the problem of using smart devices (taken by workers) to collect data in real-world applications.

Kazemi and Shahabi [18] classified the spatial crowdsourcing systems into two categories: workers' motivation and publishing models. Regarding the workers' motivation, there are two types, reward-based and self-incentivised, which inspires workers to do tasks by rewards or volunteering, respectively. Moreover, based on the publishing models, there are two modes, *worker selected tasks* (WST) [5], [13] and *server assigned tasks* (SAT) [18], [10], in which tasks are accepted by workers or assigned by workers, respectively.

Our PB-SC problem is reward-based and follows the SAT mode. Prior works on the SAT mode [18], [10] assigned existing workers to tasks in the spatial crowdsourcing system with distinct goals, for example, maximizing the number of the completed tasks on the server side [18], or the reliability-and-diversity score of assignments [10]. In contrast, our PB-SC problem in this paper has a different goal, that is, maximizing the total quality score of assignments and under the budget score. As a result, we should design specific greedy and D&C algorithms for our PB-SC problem, that maximize the quality score (under the budget constraint) rather than other metrics (e.g., the reliability-and-diversity score in [10]), which cannot directly borrow from previous works. Different from the divide-and-conquer algorithm in existing work [10] that keeps dividing each problem into two subproblems, our PB-SC_D&C utilizes a cost model to estimate a parameter g , then keeps dividing each problem into g subproblems such that the total time cost is minimized.

Most importantly, prior works on spatial crowdsourcing [18], [13], [10] usually considered the assignment strategy only over the currently available workers/tasks. In this paper, we find that this strategy might lead to local optimality, and thus propose a new prediction-based assignment strategy that estimates location/quality distributions of future workers/tasks, with which better local optimal assignments are expected. Thus, our PB-SC problem aims to design the assignment strategy over both current and predicted workers/tasks, which has not been investigated before. Therefore, we cannot directly apply previous techniques (proposed for workers/tasks without prediction) to tackle our PB-SC problem.

VIII. CONCLUSION

In this paper, we study an important spatial crowdsourcing problem, named *prediction-based spatial crowdsourcing* (PB-SC), which assigns moving workers with spatial tasks satisfying the budget constraint of the traveling cost and achieving the maximal total quality score. In order to obtain global optimal assignments, in this paper, our PB-SC problem will be based on the assignment selection strategy over both current and (predicted) future workers/tasks. In this paper, we propose an accurate prediction approach to estimate both quality/location distributions of workers/tasks. We prove that the PB-SC problem is NP-hard, and thus intractable.

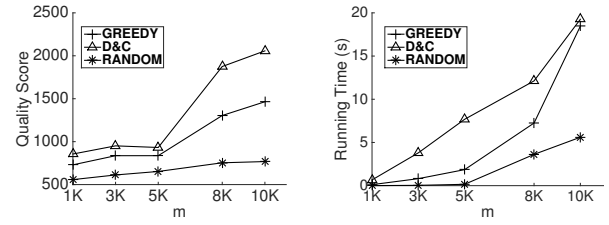


Fig. 14. Effect of the Number, m , of Tasks (Synthetic Data).

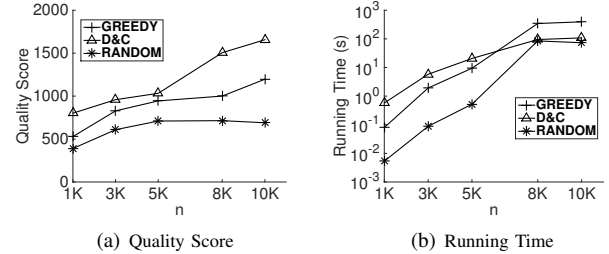


Fig. 15. Effect of the Number, n , of Workers (Synthetic Data).

Alternatively, we propose efficient approximate algorithms, including *greedy* and D&C approaches, to compute the best worker-and-task assignments, which are guided by a novel cost model. Extensive experiments have been conducted to confirm the efficiency and effectiveness of our proposed PB-SC processing approaches.

REFERENCES

- [1] Gigwalk. <http://www.gigwalk.com>.
- [2] Google street view. <https://www.google.com/maps/views/streetview>.
- [3] Taskrabit. <https://www.taskrabit.com>.
- [4] Waze. <https://www.waze.com>.
- [5] F. Alt, A. S. Shirazi, A. Schmidt, U. Kramer, and Z. Nawaz. Location-based crowdsourcing: extending crowdsourcing to the real world. In *NordiCHI 2010: Extending Boundaries*, 2010.
- [6] S. Borzsony, D. Kossman, and K. Stocker. The skyline operator. In *ICDE*, 2001.
- [7] M. F. Bulut, Y. S. Yilmaz, and M. Demirbas. Crowdsourcing location-based queries. In *PERCOM Workshops*, 2011.
- [8] Z. Chen, R. Fu, Z. Zhao, Z. Liu, L. Xia, L. Chen, P. Cheng, C. C. Cao, and Y. Tong. gmission: A general spatial crowdsourcing platform. *PVLDB*, 7(13), 2014.
- [9] P. Cheng, X. Lian, L. Chen, and C. Shahabi. Prediction-based task assignment on spatial crowdsourcing (technical report). <http://arxiv.org/abs/1512.08518>, 2015.
- [10] P. Cheng, X. Lian, Z. Chen, R. Fu, L. Chen, J. Han, and J. Zhao. Reliable diversity-based spatial crowdsourcing by moving workers. *PVLDB*, 8(10), 2015.
- [11] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *SIGKDD*, 2011.
- [12] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos. Anonymize: privacy-aware people-centric sensing. In *MobiSys*, 2008.
- [13] D. Deng, C. Shahabi, and U. Demiryurek. Maximizing the number of worker's self-selected tasks in spatial crowdsourcing. In *SIGSPATIAL GIS*, 2013.
- [14] C. M. Grinstead and J. L. Snell. *Introduction to probability*. American Mathematical Soc., 2012.
- [15] B. E. Hansen. Lecture notes on nonparametrics. *Lecture notes*, 2009.
- [16] G. Jovanovic-Dolecek. Demo program for central limit theorem. In *MWSCAS*, volume 1, 1997.
- [17] S. S. Kanhere. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In *MDM*, volume 2, 2011.
- [18] L. Kazemi and C. Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing. In *SIGSPATIAL GIS*, 2012.
- [19] S. H. Kim, Y. Lu, G. Constantinou, C. Shahabi, G. Wang, and R. Zimmermann. Mediaq: mobile multimedia management system. In *MMSys*, 2014.
- [20] C. L. Lawson and R. J. Hanson. Solving least squares problems. volume 161. SIAM, 1974.

- [21] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *ICDE*, 2012.
- [22] V. V. Vazirani. Approximation algorithms. Springer Science & Business Media, 2013.

APPENDIX

A. Proof of Lemma 2.1

Proof: We prove the lemma by a reduction from the 0-1 knapsack problem. A 0-1 knapsack problem can be described as follows: Given a set, C , of n items a_i numbered from 1 up to n , each with a weight w_i and a value v_i , along with a maximum weight capacity W , the 0-1 knapsack problem is to find a subset C' of C that maximizes $\sum_{a_i \in C'} v_i$ subjected to $\sum_{a_i \in C'} w_i \leq W$.

For a given 0-1 knapsack problem, we can transform it to an instance of PB-SC as follows: at timestamp p , we give n pairs of worker and task, such that for each pair of worker and task $\langle w_i, t_i \rangle$, the traveling cost $c_{ii} = w_i$ and the quality score $q_{ii} = v_i$. Also, we set the global budget $B = W$. At the same time, for any pair of worker and task $\langle w_i, t_j \rangle$ where $t_j \neq t_i$, we make $c_{ij} \gg c_{ii}$ and $q_{ij} \leq q_{ii}$. Thus, in the assignment result, it is only possible to select worker-and-task pairs with same subscripts. Then, for this PB-SC instance, we want to achieve an assignment instance set I_p of some pairs of worker and task with same subscripts that maximizes the quality score $\sum_{\langle w_i, t_i \rangle \in I_p} q_{ii}$ subjected to $\sum_{\langle w_i, t_i \rangle \in I_p} c_{ii} \leq B$.

Given this mapping, we can show that the 0-1 knapsack problem instance can be solved, if and only if the transformed PB-SC problem can be solved.

This way, we can reduce 0-1 knapsack problem to the PB-SC problem. Since 0-1 knapsack problem is known to be NP-hard [22], PB-SC is also NP-hard, which completes our proof. ■

B. The Pseudo Code of the Grid-based Prediction Algorithm

Figure 16 shows the pseudo code of our grid-based prediction algorithm, namely PB-SC_Prediction, which predicts the number of workers/tasks in each cell, $cell_i$, of the grid index by using the linear regression (lines 3-4), and generates worker/task samples with the estimated sizes (lines 5-6).

```

Procedure PB-SC_Prediction {
  Input: worker sets  $\mathbb{W} = \{W_{p-w+1}, \dots, W_p\}$  and task sets  $\mathbb{T} = \{T_{p-w+1}, \dots, T_p\}$  in the latest  $w$  rounds, and a grid index  $\mathcal{I}$ 
  Output: future workers and tasks in  $W_{p+1}$  and  $T_{p+1}$ , respectively, for the next round at timestamp  $(p+1)$ 
  (1) let  $W_{p+1} = \emptyset$  and  $T_{p+1} = \emptyset$ 
  (2) for each cell  $cell_i$  in  $\mathcal{I}$ 
  (3)   estimate the future number,  $|W_{p+1}^{(i)}|$ , of workers in  $cell_i$  by the linear regression
  (4)   estimate the future number,  $|T_{p+1}^{(i)}|$ , of tasks in  $cell_i$  by the linear regression
  (5)   randomly generate  $|W_{p+1}^{(i)}|$  worker samples for  $cell_i$ , and add them to  $W_{p+1}$ 
  (6)   randomly generate  $|T_{p+1}^{(i)}|$  task samples for  $cell_i$ , and add them to  $T_{p+1}$ 
  (7) return  $W_{p+1}$  and  $T_{p+1}$ 
}

```

Fig. 16. The Grid-Based Worker/Task Prediction Algorithm.

C. Cost-Model-Based Estimation of the Best Number of the Decomposed Subproblems

The Cost, F_D , of Decomposing Subproblems. From Algorithm PB-SC_Decomposition (in Figure 6), we first need to retrieve all valid worker-and-task assignment pairs (line 3), whose cost is $O(m' \cdot n')$, where m' and n' are the numbers of both current/future tasks and workers, respectively. Then, we will divide each problem into g subproblems, whose cost

is given by $O(m' \cdot g + m')$ on each level. For level k , we have m'/g^k tasks in each subproblem $M_i^{(k)}$. We will further partition it into g more subproblems, $M_j^{(k+1)}$, recursively, and each one will have m'/g^{k+1} tasks. In order to obtain m'/g^{k+1} tasks in each subproblem $M_j^{(k+1)}$, we first need to find the anchor task, which needs $O(m'/g^k)$ cost, and then retrieve the rest tasks, which needs $O(m/g^{k+1})$ cost. Moreover, since we have g^{k+1} subproblems on level $(k+1)$, the cost of decomposing tasks on level k is given by $O(m' \cdot g + m')$.

Since there are totally $\log_g(m')$ levels, the total cost of decomposing the PB-SC problem is given by:

$$F_D = m' \cdot n + (m' \cdot g + m') \cdot \log_g(m').$$

The Cost, F_C , of Recursively Conquering Subproblems.

Let function $F_C(x)$ be the total cost of recursively conquering a subproblem, which contains x spatial tasks. Then, we have the following recursive function:

$$F_C(m') = g \cdot F_C\left(\frac{m'}{g}\right).$$

Assume that deg_t is the average number of tasks in the assignment instance set I_p . Then, the base case of function $F_C(x)$ is the case that $x = 1$, in which we greedily select one worker-and-task pair from deg_t pairs to achieve the highest quality score (via the *greedy algorithm*). Thus, we have $F_C(1) = 2deg_t^2$.

From the recursive function $F_C(x)$ and its base case, we can obtain the total cost of the recursive invocation on levels from 1 to $\log_g(m')$ below:

$$\sum_{k=1}^{\log_g(m')} F_C(m'/g^k) = \frac{2(m'-1)deg_t^2}{g-1}.$$

The Cost, F_M , of Merging Subproblems. Next, we provide the cost, F_M , of merging subproblems by resolving conflicts. For level k , we have g^k subproblems and m'/g^k tasks in each subproblem $M_i^{(k)}$. When merging the result of one subproblem $M_i^{(k)}$ to the current maintained assignment instance set I_p , there are at most m'/g^k conflicted workers as each task is only assigned with one worker. In addition, for level k , we need to merge the results of $(g^k - 1)$ subproblems to the current maintained assignment instance set. Moreover, when resolving conflict of one worker, we may need to greedily pick one available worker from deg_t workers, which costs $2deg_t^2$.

Therefore, the worst-case cost of resolving conflicts during resolving conflicts of workers can be given by:

$$F_M = \sum_{k=1}^{\log_g(m')} 2deg_t^2(g^k - 1) \frac{m}{g^k} = 2deg_t^2 \left(\frac{m \log m}{\log g} - \frac{g(m-1)}{g-1} \right).$$

The Cost, F_B , of the Budget Adjustment on Subproblems.

Then, we provide the cost, F_B , of the budget adjustment in line 15 of the D&C algorithm (in Figure 8), or lines 17-28 in procedure PB-SC_Budget_Constrained_Selection (Figure 7). Since each task is assigned with at most one worker, the number of candidate pairs is at most same as the number of tasks. For a subproblem with m'/g^k tasks on level k , lines 20-25 of Figure 7 need at most $(m'/g^k)^2$ cost. In addition, line 26 needs $|S_p|$ cost, which is also at most $(m'/g^k)^2$ cost. There are at most (m'/g^k) iterations, each of which selects at least one pair.

Therefore, the cost of the budget adjustment while merging subproblems can be given by:

$$F_B = \sum_{k=0}^{\log_g(m')} g^k \cdot 2\left(\frac{m'}{g^k}\right)^3 = \sum_{k=0}^{\log_g(m')} \frac{2m'^2}{g^{2k}} = \frac{2g^2(m'^2 - 1)}{g^2 - 1}. \quad (11)$$

The Total Cost of the g -D&C Approach. The total cost, $cost_{D\&C}$, of the D&C algorithm can be given by summing up four costs, F_D , F_C , F_M , and F_B . That is, we have:

$$cost_{D\&C} = F_D + \sum_{k=1}^{\log_g(m')} F_C(m'/g^k) + F_M + F_B. \quad (12)$$

We take the derivation of $cost_{D\&C}$ (given in Eq. (12)) over g , and let it be 0. In particular, we have:

$$\begin{aligned} \frac{\partial cost_{D\&C}}{\partial g} &= \frac{m' \log m' (g \log g - g - 1 - 2deg_t^2)}{g \log^2(g)} \\ &\quad - \frac{4g(m'^2 - 1)}{(g^2 - 1)^2} = 0. \end{aligned} \quad (13)$$

We notice that when $g = 2$, $\frac{\partial cost_{D\&C}}{\partial g}$ is much smaller than 0 but increases quickly when g grows. In addition, g can only be an integer. Then we can try the integers like 2, 3, and so on, until $\frac{\partial cost_{D\&C}}{\partial g}$ is above 0.

D. Results with Different Worker-Task Distributions

In this section, we present the experimental results for workers and tasks with different location distributions, where parameters of synthetic data are set to default values. We denote the Uniform distribution as U, the Gaussian distribution as G, and the Zipf distribution as Z. Then, for $\langle worker-task \rangle$ distributions, we tested the quality score and running time over 9 distribution combinations, including G-U, G-G, G-Z, U-U, U-G, U-Z, Z-U, Z-G, and Z-Z, and the results are shown in Figures 17 and 18.

Similar to previous results, as shown in Figure 17, the D&C algorithm can achieve the highest quality score, compared with GREEDY and RANDOM, over all the 9 worker/task distribution combinations. For the running time, as illustrated in Figure 18, with different combinations of worker/task distributions, D&C can achieve low time cost in most cases. Only for Z-U and Z-G, D&C incurs higher time cost than GREEDY and RANDOM, due to the unbalanced distributions of workers and tasks. In particular, GREEDY and RANDOM iteratively assign one valid pair and maintain the rest of valid pairs in each iteration. When the distributions of workers and tasks are similar, for example, G-G, U-U, and Z-Z, running times of GREEDY and RANDOM become longer than D&C. Especially, for Z-Z, almost all the workers can reach all the tasks, which leads to the highest number of valid pairs among all the 9 distribution combinations. As a result, both GREEDY and RANDOM need much higher running time than that of other distribution combinations. In general, with different worker and task distributions, our GREEDY and D&C can both achieve high quality scores (with small time cost).

E. The PB-SC Performance vs. the Number, R , of Rounds and the Unit Price C w.r.t. Distance $dist(w_i, t_j)$

The PB-SC Performance vs. the Number, R , of Rounds. Figure 19 reports the experimental results for different numbers, R , of rounds from 10 to 25 on synthetic data sets, where other parameters are set to default values. In Figure 19(a), when the number, R , of rounds increases, the total

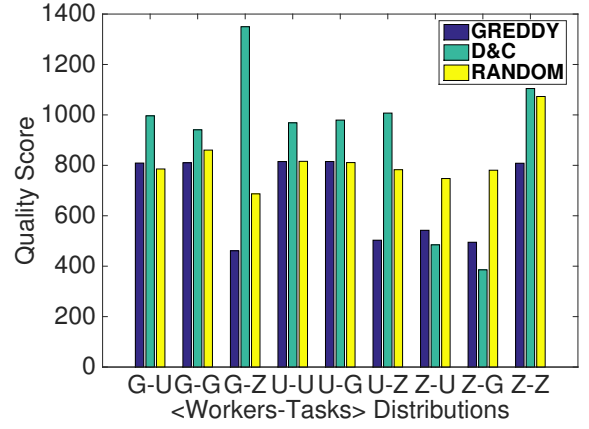


Fig. 17. Effect of Distributions of Workers and Tasks on Quality Score (Synthetic Data).

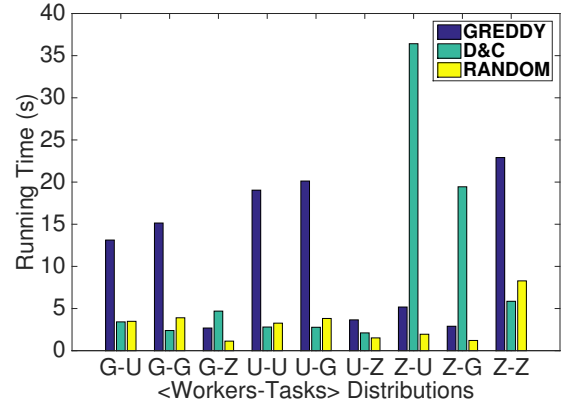


Fig. 18. Effect of Distributions of Workers and Tasks on Running Time (Synthetic Data).

quality score of three PB-SC approaches also increases. Since we consider a fixed time interval P with more rounds (each with budget B), the total quality score within interval P expects to increase for more rounds. D&C can achieve higher quality scores than GREEDY.

In Figure 19(b), when R becomes larger, the running time of all the three tested approaches decreases. This is because, given m tasks and n workers within time interval P , for more rounds, the average number of workers/tasks per round decreases, which leads to lower time cost per round. Similar to previous results, the running time of GREEDY is lower than that of D&C.

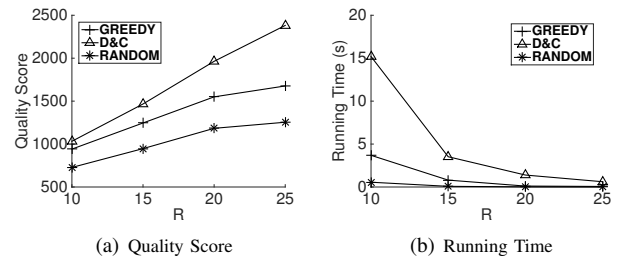


Fig. 19. Effect of the Number, R , of Rounds (Synthetic Data).

The PB-SC Performance vs. the Unit Price C w.r.t. Distance $dist(w_i, t_j)$. Figure 20 illustrates the experimental results on different unit prices C w.r.t. distance $dist(w_i, t_j)$ from 5 to 20 over synthetic data, where other parameters are set to default values. In Figure 20(a), when the unit price C increases, the total quality scores of all the three approaches decrease. This is because for larger C , the number of valid worker-and-task pairs for each round decreases, under the budget constraint. Thus, the total quality score of all the selected assignments also expects to decrease for large C . Similar to previous results, D&C has higher quality scores than GREEDY. In Figure 20(b), running times of GREEDY and RANDOM are not very sensitive to C . However, with large C , the running time of D&C increases, since we need to check the constraint of budget from lower divide-and-conquer levels, which increases the total running time. For different C values, GREEDY has lower running times than D&C.

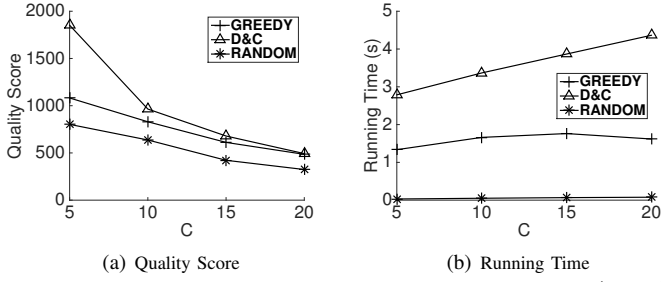


Fig. 20. Effect of the Unit Price C w.r.t. Distance $dist(w_i, t_j)$ (Synthetic Data).